

Maria Cristina Bassi

Formati MARC-XML e servizi derivati: resoconto sulla sessione al congresso IFLA 2005

XML is not an immediate business of library community, but in fact we can see now this interest at the highest level of library community – at the World Library and Information Congress.
(Vladimir Skvortsov, intervento alla sessione “MARC/XML derivatives” dell’IFLA 2005)

Then the serpent saith unto Adam, “Lo, why doth thy catalogue service not use XML?” And Adam saith, “Verily, Z39.50 worketh just fine.” But the serpent, who was subtle of tongue, saith unto him, “But XML is more fashionable.” And, behold, Adam was deceived, and did fall.
(Mike Taylor, presentazione alla sessione “MARC/XML derivatives” dell’IFLA 2005)

Da qualche anno nel campo dell’automazione di biblioteca si parla molto dell’uso del metalinguaggio XML per rappresentare i record catalografici (o – più in generale – i metadati).

L’eXtensible Markup Language applicato ai formati MARC (Machine-Readable Cataloging) è stato oggetto di alcune presentazioni durante la WLIC World Library and Information Congress 2005 svoltasi ad Oslo in agosto.

In particolare, le sezioni UNIMARC e Information Technology dell’IFLA hanno organizzato la sessione “MARC/XML derivatives: the state of the art”, di taglio decisamente tecnico ma di grande interesse a parere di chi scrive. I 4 interventi alla sessione hanno fornito – in sequenza – lo stato dell’arte sul Manuale UNIMARC in XML, sugli schemi di metadati MARCXML, MarcXchange e UNIMARC/XML, sui protocolli Z39.50, SRW e SRU.

Per conformità allo sviluppo ‘cronologico’ delle applicazioni XML citate, conviene comunque cominciare dalla relazione di **Sally McCallum** (Library of Congress, Washington D.C., USA), dal titolo *MARC XML sampler*.

Nel 2002 la Library of Congress ha adottato uno schema XML per il formato MARC21 – chiamato MARCXML *slim* (magro) in quanto si tratta di uno standard molto breve – che fornisce un equivalente esatto del record MARC21 (ISO2709) rendendo possibile una conversione (in entrambi i sensi) senza perdita di dati. Alla base dell’attuale processo di standardizzazione presso l’ISO di una versione XML per il formato MARC (MarcXchange), MARCXML può essere usato per lo scambio di dati o come ‘mezzo di trasporto’ per ulteriori trasformazioni/elaborazioni. L’insieme degli strumenti disponibili (MARC Tool Kit) è scaricabile liberamente all’indirizzo <http://www.loc.gov/marcxml> e comprende sia convertitori per trasformare record da MARC21 (ISO2709) a MARCXML e viceversa, sia varie altre interessanti trasformazioni. In particolare si citano:

- trasformazioni del formato in DC (Dublin Core) e MODS (Metadata Object Description Schema),
- trasformazioni del formato da DC, MODS, ONIX (Online Information Exchange) e oai_marc,
- visualizzazione da MARCXML a FRBR (Functional Requirements for Bibliographic Records),
- validazione di record MARCXML,

- trasformazioni alla e dalla precedente DTD (Document Type Definition) XML MARC21.

Tali trasformazioni, osserva la relatrice, sono considerate preziose dalla Library of Congress in quanto aiutano a mantenere le economie e l'interoperabilità raggiunte grazie all'uso di un formato comune. Esistono numerose applicazioni specializzate in cui MARCXML è impiegato per comunicare record MARC e per utilizzare i molti strumenti XML che permettono di intervenire frequentemente sui record trattandoli in modo più agevole rispetto all'ISO2709:

- *Switch* di metadati. Il progetto dei servizi di terminologia di OCLC – relativo ai vocabolari di organizzazione della conoscenza (authority file, sistemi di intestazioni di soggetto, thesauri e schemi di classificazione) – usa la versione MARCXML del MARC21 che permette di accrescere i vocabolari con le utilities XML e i fogli di stile XSLT (eXtensible Stylesheet Language Transformations), rendendo più semplice combinare dati non-MARC con dati MARC e altri dati formattati in XML;
- Editing di record e applicazioni crosswalk. Alcuni strumenti di mappatura che utilizzano MARCXML permettono conversioni di dati da Dublin Core, Encoded Archival Description (EAD), e metadati di oggetti geospaziali (formato FGDC, Federal Geographic Data Committee) a MARC21. In tal modo facilmente si possono prendere i metadati DSpace creati per un *repository* DSpace e convertirli in MARC21 via MARCXML al fine di integrarli in un catalogo di biblioteca;
- Manipolazione di record e strumento di gestione. Per esempio con lo scopo di identificare speciali gruppi di record catalografici, possono essere usati strumenti XML per elaborare in XML blocchi di record dal catalogo;
- Protocolli basati su XML. MARCXML è largamente impiegato per scambiare dati in MARC21 da parte dei nuovi protocolli in ambiente XML, e cioè OAI-PMH (Open Archive Initiative - Protocol for Metadata Harvesting) e SRW/SRU (per l'Information Retrieval). Ad esempio: tramite un foglio di stile XSL – parte del MARC Tool kit – è possibile convertire record oai_marc in MARCXML (supportato dal protocollo OAI); i dati del progetto VIAF (Virtual International Authority File) saranno 'visti' dal protocollo SRU come metadati MARCXML.

In definitiva – afferma McCallum – MARCXML costituisce la base per evoluzioni future mantenendo allo stesso tempo i vantaggi della standardizzazione.

Joaquim Ramos De Carvalho (Università di Coimbra, Coimbra, Portogallo) ha invece discusso di *An XML representation of the UNIMARC Manual: a working prototype*.

Si può pensare ad un approccio più articolato rispetto a quello della Library of Congress, osserva il relatore: cioè non solo un semplice schema per il trasporto (come MARCXML *slim*), ma anche un altro schema per rappresentare le regole che definiscono un record UNIMARC valido (=corretto). Nel 2004 la società BookMARC è stata incaricata dalla Biblioteca Nazionale portoghese – nel contesto della *core activity* IFLA sull'UNIMARC – di sviluppare un prototipo di schema XML che potesse rappresentare il contenuto del Manuale UNIMARC Bibliographic. Il Manuale UNIMARC/B comprende infatti vari tipi di informazioni che devono essere incorporate nei sistemi di automazione per biblioteche: regole strutturali (di che parti si compone un record bibliografico), regole di contenuto (i campi richiesti e i valori accettati per varie componenti del record), esempi, note esplicative, rimandi ad altri punti del manuale. I software per biblioteche dovrebbero poter usare tali dati per scopi quali la validazione di record, la produzione di indici, la formattazione di record per visualizzarli e la predisposizione di aiuti leggibili ad occhio nudo. Era quindi necessario che la versione XML del Manuale UNIMARC:

- supportasse la validazione automatica dei record,
- trasportasse i record bibliografici in termini *human-readable*,
- fornisse i contenuti del Manuale in formati diversi.

Il risultato del lavoro di BookMARC è disponibile a <http://www.bookmarc.pt/unimarc>: si tratta di uno schema per il Manuale UNIMARC e di un set di strumenti software che spiegano come usarlo (compresi fogli di stile per formattare il manuale in HTML, per validare i record e per formattarli in HTML). Tale prototipo XML del Manuale UNIMARC dimostra che è possibile sia usare l'XML per

rappresentare le regole semantiche del formato UNIMARC, sia lavorare in modo raffinato sui record tramite mezzi standard quali le trasformazioni XSLT. Il vantaggio maggiore sta infatti nelle possibilità create dall'elaborazione dei record grazie alla sofisticate capacità dei fogli di stile XSLT, che potrebbero essere prodotti automaticamente e con facilità per scopi di validazione e formattazione. Inoltre Carvalho ipotizza ulteriori applicazioni del prototipo, tra cui l'aggiornamento automatico dei sistemi software importando una nuova versione del Manuale [questa possibilità è interessante se si pensa ad es. alla recente pubblicazione dell'Update 5 del formato UNIMARC/Bibliographic, aggiornamento che per forza di cose non è integrato in release già esistenti di software di automazione per biblioteche]. Tra le questioni aperte vengono citate le regole condizionali (aspetto complesso da rendere in XML), la 'neutralità' dello schema del Manuale relativamente al formato XML per i record bibliografici (per gli esempi si è usato il formato MARCXML *slim*, in quanto al momento non esiste uno standard per rappresentare l'UNIMARC in XML), i formati non bibliografici della famiglia UNIMARC (bisognerebbe testare se lo schema così com'è riesce a gestire questi altri tipi di record e le loro regole) e infine l'aspetto delle licenze e del copyright.

Ha quindi preso la parola **Vladimir Skvortsov**, che ha parlato anche in rappresentanza di Olga Zhlobinskaya e Alla Pashkova (Biblioteca Nazionale di Russia, S. Pietroburgo, Federazione russa): il loro paper si è focalizzato su ***UNIMARC/XML Slim Schema: living in a new environment***. Dato l'inglese traballante di Skvortsov, la presenza di diapositive è stata essenziale per poter seguire il discorso.

Il ruolo dell'XML nei sistemi di automazione di biblioteca – con riguardo alle funzioni di base dei formati bibliografici di tipo MARC (immagazzinamento e trasporto dei dati) – si profila come estremamente rilevante. 1) Immagazzinamento. Possiamo considerare XML e relative tecniche come un normale database (cioè un DBMS, Data Base Management System); inoltre i documenti XML possono essere immagazzinati nei cosiddetti 'database XML nativi', con vantaggi in termini di velocità d'accesso. 2) Trasporto. Tuttora i record in formato MARC vengono trasportati principalmente attraverso il 'contenitore' ISO2709. Rispetto ad ISO2709 XML presenta numerosi vantaggi, che ne fanno uno standard de facto per lo scambio di dati tramite interfaccia web:

- XML non ha alcune limitazioni dell'ISO2709 (per esempio sulla lunghezza del record);
- i documenti XML sono leggibili ad occhio nudo;
- i documenti XML – nonostante la loro relativa facilità di creazione ed elaborazione – permettono di produrre informazioni strutturate che potrebbero essere facilmente 'comprese' da un computer;
- la sintassi XML – piuttosto stringente – dà la possibilità di formalizzare il controllo in fase di creazione del documento;
- XML supporta le relazioni gerarchiche, permettendo quindi di rispecchiare la struttura gerarchica dei record bibliografici;
- visualizzati o stampati, i documenti XML possono includere informazioni sul layout dei dati che il linguaggio per la descrizione dello stile (XSL) è in grado di gestire;
- fin dall'origine XML è integrato con le interfacce web, e di conseguenza con i browser web.

Un esempio di set di strumenti suggerito da Skvortsov, Zhlobinskaya e Pashkova per l'immagazzinamento e il trasporto di dati bibliografici in ambiente XML comprende: a) uno Schema XML *slim* (schema per il trasporto) che definisca i formati MARC al livello più comune, giocando in tal modo il ruolo dell'ISO2709; tale schema dovrebbe essere semplice, garantire una conversione tra XML e ISO2709 – e viceversa – senza perdita di dati, e risultare stabile rispetto a qualsiasi cambiamento nei formati MARC; b) uno Schema normativo XML (il formato MARC convertito in XML); infine c) un kit di

- utilities per la conversione da ISO2709 a XML e da XML a ISO2709;
- strumenti per validare documenti XML relativamente alla conformità allo schema *slim*;
- strumenti per validare documenti XML relativamente alla conformità allo schema normativo;
- fogli di stile per la visualizzazione dei record in modalità user-friendly (etichetta - contenuto);
- strumenti per validare record bibliografici impiegando trasformazioni XSLT.

Attualmente lo schema *slim* più importante è il MARC21XML della Library of Congress, che però dà problemi se applicato all'UNIMARC (in particolare, non sono previsti i campi *embedded* e una parte dei valori degli indicatori UNIMARC). Sviluppando uno schema XML *slim* per UNIMARC, Skvortsov, Zhlobinskaya e Pashkova si sono comunque basati su MARC21XML, con l'obiettivo – però – di creare uno schema XML che fosse adatto a qualunque formato MARC. Tale schema è stato realizzato dal Centre for Computer Technologies of the Ural State University di Ekaterinburg in collaborazione col National Service on RUSMARC format development. I requisiti su cui si è incardinata la progettazione dello schema sono: perfetta conversione in documenti XML dei record MARC in ISO2709; validazione dei record (controllo della correttezza della struttura del record); considerazione di specifiche caratteristiche del formato UNIMARC (si vedano gli *embedded fields* e la gamma completa dei valori degli indicatori); maggiore indipendenza possibile da qualsiasi particolare 'dialetto' del MARC; stabilità nei confronti dei cambiamenti del formato UNIMARC. Lo Schema XML *slim* si può trovare a http://www.rba.ru/rusmarc/soft/rusmarc_slim.xsd.

L'ultimo intervento della sessione è stato quello di Mike Taylor (Index Data, Londra, UK), che ha presentato *Delivering MARC/XML records from the Library of Congress catalogue using the open protocols SRW/U and Z39.50* (paper scritto assieme al collega Adam Dickmeiss della sede di Copenhagen di Index Data). Le slides illustrate da Taylor – vivacizzate da battute e ironia – hanno piacevolmente colpito l'uditorio presente alla sessione di Oslo... Tali diapositive sono fruibili dai bibliotecari, che non sempre possono vantare un grosso background tecnico.

Grazie all'uso generalizzato del formato MARC per la rappresentazione dei dati catalografici e dell'ISO2709 per il loro trasporto, le biblioteche hanno raggiunto un grado di cooperazione e interoperabilità che in altri ambiti (ad es. nella gestione delle collezioni nei musei) si è ben lungi dall'ottenere. In realtà – osservano Mike Taylor e Adam Dickmeiss – spesso sottovalutiamo l'importanza di altri standard: quelli che servono a cercare-localizzare e recuperare i record catalografici sulle reti. Il protocollo Z39.50 ha avuto in questo senso un ruolo dominante nell'ultimo decennio. Con il successo dell'eXtensible Markup Language, analogamente allo spostamento dagli storici formati binari di rappresentazione dei dati a quelli di tipo XML (da MARC21 a MARCXML e MarcXchange), per le funzioni di *search-and-retrieve* stiamo passando dai protocolli binari a quelli basati su XML. Nel 2000 si è infatti cominciato a riscrivere lo standard Z39.50 al fine di renderlo più adatto al nuovo ambiente informativo: il risultato è costituito da due protocolli – SRW (Search/Retrieve Web-service) e SRU (Search/Retrieve URL) – che vengono mantenuti da un comitato editoriale composto da compagnie commerciali, biblioteche e università. SRW e SRU sono equivalenti, ma SRU ha preso piede maggiormente in quanto impiega un meccanismo più semplice (diversamente da SRW, per il trasporto non usa SOAP Simple Object Access Protocol ma invece va a prendere URL 'ricche'). Entrambi si basano sul protocollo HTTP (Hypertext Transfer Protocol), di cui costituiscono un notevole raffinamento per quanto riguarda l'Information Retrieval. In SRW/SRU è stato migliorato il linguaggio d'interrogazione strutturato che Z39.50 usa per esprimere query 'ricche': tale formato di query testuale è conosciuto come CQL, Common Query Language. Taylor e Dickmeiss riferiscono poi del lavoro con SRW/SRU eseguito dalla società Index Data per la Library of Congress. Quando le tecnologie cambiano, durante la fase di transizione e di sovrapposizione di due standard i *service providers* devono garantire contemporaneamente due tipi di servizi per esigenze di flessibilità. Nel 2003 la Library of Congress – decidendo di trasportare i propri record MARCXML via SRW così come i record in MARC21 su Z39.50 – si è trovata di fronte al problema di decidere quale approccio seguire. La soluzione messa in atto da Index Data per la LOC prevede un gateway che funziona da server SRW/MARCXML e soddisfa le richieste a lui mandate funzionando anche come client Z39.50/MARC21. Index Data ha quindi fornito un'entrata da SRW-a-Z39.50 che sta 'di fronte' al server Endeavor Z39.50 della Library of Congress. Tale gateway – rilasciato sotto licenza open source (GNU General Public License) – è ora liberamente disponibile a <http://indexdata.com/yazproxy>. L'entrata SRW è un'insolita applicazione dello Z39.50, dato che quest'ultimo è stato primariamente inteso per la comunicazione tra client remoti e server (nel caso in oggetto, invece, il client e il server Z39.50 si trovano su macchine che come minimo rientrano nella stessa LAN – Local Area Network). Il gateway SRW-a-Z39.50 di Index Data si chiama "Yaz [Yet

Another Z39.50 toolkit] Proxy” e presenta alcuni vantaggi collaterali tra cui la riduzione del carico sul server e il miglioramento della performance del client. Può essere usato per stare ‘davanti’ a qualsiasi server Z39-50 e ne esistono diverse installazioni a livello mondiale. Yaz Proxy eroga record MARCXML via Z39.50, convertendo al volo i record MARC21 originari; inoltre – usando fogli di stile XSLT – il gateway può essere configurato per trasformare ulteriormente i documenti MARCXML derivanti verso altri formati XML, compresi Dublin Core, MODS, METS (Metadata Encoding and Transmission Standard) e oai_dc. Una critica che si può muovere a Taylor riguarda la discrepanza che si nota tra il paper in formato pdf e la presentazione in ppt a proposito del lavoro svolto da Index Data per la Library of Congress: nel primo si parla di ‘entrata’ SRW, mentre nella seconda di SRU... per cui non risulta molto chiaro quale dei due protocolli di nuova generazione sia in uso alla LOC (sembra che in realtà si tratti di entrambi).

Infine qualche parola su **MarcXchange**, citato da varie relazioni alla sessione “MARC/XML derivatives”. Come si è accennato, attualmente all’ISO viene esaminato tale schema XML – proposto nel 2004 da Danish Standards – il cui rapporto con MARCXML è quasi identico rispetto a quello che ISO2709 ha con MARC21. McCallum riassume i principali punti in cui MarcXchange differisce da ISO2709, tra cui l’aggiunta di dati a livello ‘alto’ per identificare sia la specifica designazione di contenuto MARC (esempi di valore dell’attributo possono essere MARC21, UNIMARC, Ibermarc, Danmarc...), sia il genere del record (come valori per MARC21 si indicano *bibliographic*, *authority*, *holdings*, *classification*, e *community*). MarcXchange – ricordano Taylor e Dickmeiss – è un superset di MARCXML, quindi ogni valido documento MARCXML è anche un valido documento MarcXchange con lo stesso significato. Forse, come ipotizza Taylor, sarebbe stato meglio procedere tramite ‘correzione’ di MARCXML. Per maggiori informazioni su questo schema si veda a <http://www.niso.org/international/SC4/n569.pdf>.

In conclusione la sessione è apparsa omogenea, con interventi di alto livello, anche se resta l’impressione che un maggior coordinamento tra alcuni progetti presentati sarebbe stato auspicabile: ad esempio la relazione dei russi non fa cenno al lavoro su MarcXchange, mentre quella sul Manuale UNIMARC in XML non cita lo schema UNIMARC XML del National Service on RUSMARC format development.

L’augurio è che anche i servizi di automazione di biblioteca del nostro paese abbiano la possibilità di integrare in modo efficace strumenti di conversione in XML dei record catalografici ed arrivino ad implementare i protocolli SRW/U non solo a livello sperimentale.

Ad ogni modo, un’occasione per fare il punto su formati MARC/XML e protocolli di *search and retrieve* sarà probabilmente la conferenza internazionale “Unimarc & friends: charting the new landscape of library standards” (Lisbona, 20-21 marzo 2006), che prevede di esplorare temi quali schemi di metadati e standard di trasmissione basati sul Web.

Maria Cristina Bassi, Servizio Automazione Biblioteche - Università di Verona, e-mail: mariacristina.bassi@univr.it

Bibliografia essenziale

Joaquim Carvalho, *An XML representation of the UNIMARC Manual: a working prototype*, intervento per il World Library and Information Congress: 71th IFLA General Conference and Council. "Libraries - A voyage of discovery", August 14th - 18th 2005, Oslo, Norway, <http://www.ifla.org/IV/ifla71/papers/199e-Carvalho.pdf>.

IFLA Information Technology Section NEWSLETTER, July 2005, http://www.ifla.org/VII/s21/news/ITSnewsletter_July_05.pdf.

Sally H. McCallum, *MARCXML Sampler*, intervento per il World Library and Information Congress: 71th IFLA General Conference and Council. "Libraries - A voyage of discovery", August 14th - 18th 2005, Oslo, Norway, <http://www.ifla.org/IV/ifla71/papers/175e-McCallum.pdf> .

Sally McCallum, *Library of Congress*, in *ICABS Activity Report August 2004 - January 2005*, <http://www.ifla.org/VI/7/annual/ICABS-Report01-2005.pdf> .

Sally McCallum, *Library of Congress*, in *ICABS Activity Report February - July 2005*, <http://www.ifla.org/VI/7/annual/ICABS-Report02-2005.pdf> .

Vladimir Skvortsov, Olga Zhlobinskaya e Alla Pashkova, *UNIMARC XML Slim Schema: Living in new environment*, intervento per il World Library and Information Congress: 71th IFLA General Conference and Council. "Libraries - A voyage of discovery", August 14th - 18th 2005, Oslo, Norway, <http://www.ifla.org/IV/ifla71/papers/064e-Skvortsov.pdf> .

Mike Taylor, *Delivering MARC/XML records from the Library of Congress catalogue using the open protocols SRW/U and Z39.50*, presentazione per il World Library and Information Congress: 71th IFLA General Conference and Council. "Libraries - A voyage of discovery", August 14th - 18th 2005, Oslo, Norway, <http://www.indexdata.dk/yazproxy/ifla2005/marcxml-proxy.ppt> .

Mike Taylor e Adam Dickmeiss, *Delivering MARC/XML records from the Library of Congress catalogue using the open protocols SRW/U and Z39.50*, intervento per il World Library and Information Congress: 71th IFLA General Conference and Council. "Libraries - A voyage of discovery", August 14th - 18th 2005, Oslo, Norway, http://www.ifla.org/IV/ifla71/papers/065e-Taylor_Dickmeiss.pdf .